International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 103-106
e-ISSN:2278-621X

ANALYSING AND DETECTING SPAM MAILS USING ALGORITHMS

NamrathaSharath[1], Pallavi[2] & Prof Santhosh Rebello[3]

Abstract- Over the past few decades web has changed dramatically, it's interesting to compare how much Electronic-mail (email) alone has changed the way we communicate. The adoption of email as a way to communicate in business has forever reshaped how we do our work, sending and receiving information may have started from slow, uncertain process to lightning fast. Though the email segmentation is fantastic, an average internet user spends almost 15 minutes of his time on deleting from the inbox dozens of e-mails with absolutely useless texts, pictures etc. A person might define spam as citing offensive or fraudulent email solicitations that could cause him his time and could affect him too. However, it remains hard and sometimes essential to distinguish spam sent by malicious spammers from legitimate mail. The greater area of concern is phishing using mail that use spam, fake websites constructed to look identical to real sites, email and instant messages to trick you into divulging sensitive information, like bank account passwords and credit card numbers. Email filtering can be approached in two different ways, like Knowledge engineering and machine learning. Hence in this paper, we would target to search efficient methods that involves both the knowledge engineering approach that uses a set of rules has to be specified according to which emails are categorized as spam and also Machine learning approach which does not require specifying any rules.
Keywords- Phishing, Knowledge engineering, Machine Learning

1. INTRODUCTION

E-mail are messages that may have texts, file or images or any other attachments send through network to one person or group of people over telecommunication.There are different types of spam like Usenet spam, Texting spam, Mobile spam and email Spam but widely known one is Email Spam. Detecting Spam Email can also aid in avoiding unwanted phishing activities. Email spam focuses on mainly on one person. Spammers use many different methods for seeking lists of webmail or search the web for email addresses.The main reason that the spam mails are sent is to profit the individual or group that sends the spam in form of money. When spam mails are sent, they are sent to not just one, but also to millions of other people. Even if they get a miniscule amount of people to respond to their message to buy, they gain profits, also known as phishing. It is important to understand that, all undesirable emails are not spammed messages. Rather we can categorize spam emails as unsolicited commercial advertisements sent through mails over the Internet. This problem isn't new, moreover I has been increasing over the years. There is no proper convenient method yet which can detect al spam mails because they are disguised as legitimate using subject or message along with it. However spam may contain links to Internet games, or pornographic texts and pictures[1].Spam is prevalent on the Internet because the transaction cost of electronic communications is radically less than any alternate form of communication. Many ways of fighting spam have been proposed. In knowledge engineering approach uses a set of rules has to be specified according to which emails classified as spam. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering like Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbour, Rough sets and the artificial immune system[2].

2. MACHINE LEARNING IN E-MAIL CLASSIFICATION

Machine learning is a type of AI (artificial intelligence) that provides computers with the ability to learn through test data– without being explicitly programmed to classify and work. Machine learning techniques are nowadays used to automatically filter the spam e-mail in a very successful rate. Machine learning field is a subfield from the broad field of artificial intelligence, this aims to make machines able to learn like human.Machine learning is typically associated with computational statistics, data modelling, and prediction-making that can help reduce major tasks to simpler ones. In unsupervised learning we need to uncover hidden regularities (clusters) or detect anomalies in the data like spam messages[2]

E-mail classification tasks are often divided into several sub-tasks such as, Data collection and representation are mostly problem specific, and then, e-mail feature selection and feature reduction attempt to reduce the dimensionality for the remaining steps of the task

Most of the machine learning algorithms that can be used for spam detection are categorized as supervised machine learning. That is where an algorithm tries to compare inputs to desired outputs using a specific function.

[1] Department of Information Technology, AIMIT, Mangalore, Karnataka India
[2] Department of Information Technology, AIMIT, Mangalore, Karnataka India
[3] Dean, Department of IT, AIMIT, Mangalore, Karnataka India

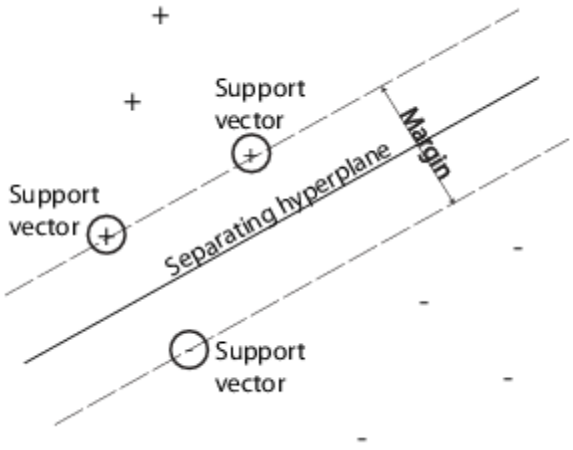
3. IDENTIFYING AND COMPARING SOLUTIONS.

In today's world one of the major issues that we have to deal with on web includes problems created through spam. These spams can also degrade the working of the system and its accuracy. To detect spam there have been many researches going on in this area, to identify different typesof spam classifiers, that help us to identify and distinguish between spam and non-spam mails. In the case of spam classification, a classifier will try to classify an email to spam or legitimate by learning certain features in the email.

3.1 Broadly opted solution:

Many different spam detection and filtering techniques popular today are: Decision tree classifier, Negative Selection Algorithm, Genetic Algorithm Support Vector Machine(SVM) Classifier, Bayesian Classifier etc. SVM algorithm is broadly opted filtering technique that offers a great volume of simplicity to achieve goals of email spam classification[4].

But there are few problems with SVM classifier because it is only useful when information can be divided into two classes. It is done by finding an ideal hyperplane that isolates every information purposes of one class from those of alternate class. In a given T element training set {(x i ,yi):x i∈RD ,y i ∈{-1,1},i=1,…,T} a linear SVM classifier: ( ) i i ii i i i i Fx a = ∑ ∑ y xx b x a · +=· +=·+ y x b x w b, Where ai ≥0, b is a bias term and D is the dimension of the input space; · is a dot-product operator, and w is the normal vector of the classification hyperplane[4].



3.2 Understanding the working of SVM:

The optimal hyperplane is found such as to maximize the classification margin, given by 2/‖w‖2, where w denotes the normal vector of the hyperplane. Points that are misclassified or lie inside the classifier's margin are identified[4].

SVM Spam Filter:

1) Initialize

Feed-in SVM classifier with a few examples of each class (spam and ham)

Train an initial SVM filters

2) Classification

- Classify xi
- Query the true label of xi
- If the filter's prediction is wrong, retrain SVM filters based on SVi + xi

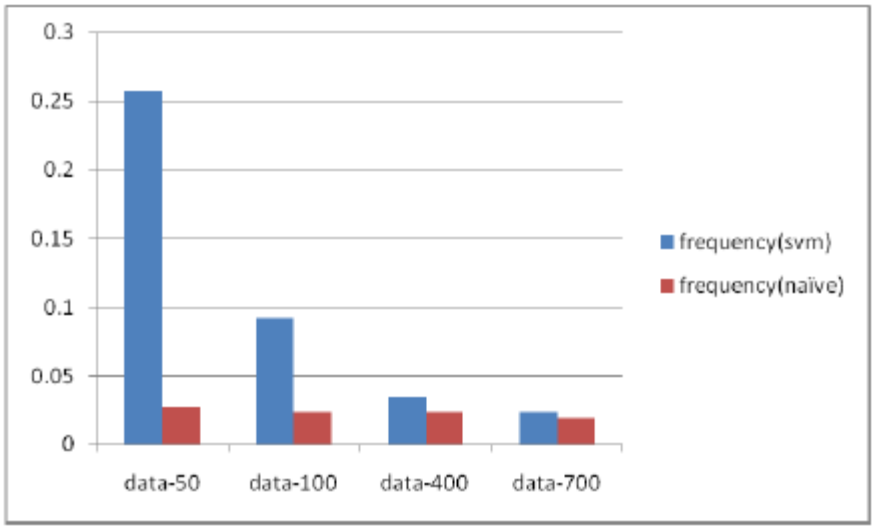
3) Complete

Repeat until xn is completed

3.3 Suggested methodology:

But there is an alternative method known as Naïve Baysian method which is very popular and open source spam detection mechanism hence we can implement it with linear computational complexity, the accuracy is way better and is less complex as compared to other ideologies[4].

To understand the working better let us compare the error rate of words which are wrongly classified in both SVM classifier as well as in Naïve Bayes[5].



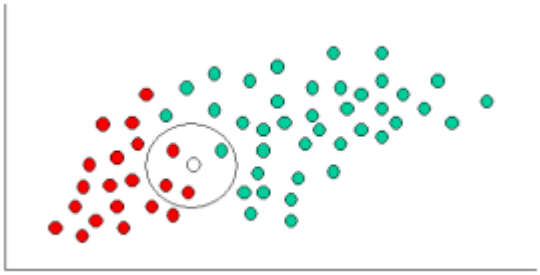
Error rate of words [8]

The graph showsclearly SVM has a greater error rate than naïve Bayes because the blue bar is higher at all instances than the red bar[8].

Hence we opt for Naïve Bayes classifier over the others.

4. NAIVE BAYESIAN CLASSIFIERS

It was in 1998 the Naïve Bayes classifier was proposed for spam recognition. Naive Bayesian Classifiers are highly scalable. Training of the large data simple can be easily done with Naive Bayesian Classifier, which takes a very less time as compared to other classifier.



Using training set for Naïve Bayes[6].

Taking a gander at the likelihood and the event of dataset existing in database with interesting information we can order information and distinguish distinctive sorts of information, which can help in recognizing ham or spam sends utilizing mix of gullible Bayes classifier with unpleasant sets and likelihood theory[7].

Bayesian classifier is taking a shot at the reliant occasions and the likelihood of an occasion happening later on that can be distinguished from the past happening of a similar occasion. This strategy can be utilized to order spam messages; words probabilities play the primary run here. On the off chance that a few words happen frequently in spam however not in ham, at that point this approaching email is most likely spam[8] . Guileless Bayes classifier system has turned into an exceptionally well known technique in mail sifting. Bayesian channel ought to be prepared to work successfully. Each word has certain likelihood of happening in spam or ham email in its database. On the off chance that the aggregate of words probabilities surpasses a specific breaking point, the channel will stamp the email to either classification.

4.1 Working of Naive Bayesian Classifiers:

The measurement we are for the most part intrigued for a token T is its spamminess (spam rating), ascertained as takes after:

$$S\,[T] = \frac{C_{Spam}(T)}{C_{Spam}(T) + C_{Ham}(T)}$$

Where CSpam(T) and CHam(T) are the quantity of spam or ham messages containing token T, individually. To compute the likelihood for a message M with tokens {T1,.....,TN}, one needs to join the individual token's spamminess to assess the

general message spamminess. A basic approach to make orders is to ascertain the result of individual token's spamminess and contrast it and the result of individual token's hamminess.

$$(H[M] = \prod_{i=1}^{N} (1 - S[T_i]))$$

Themessage is considered spam if the general spamminess item S[M] is bigger than the hamminess item H[M][9]

4.2 Naive Bayesian Classifiers Algorithm:

Stage1. Train
Parse each email into its constituent tokens. Generate a probability for each token W
$S[W] = C_{spam}(W) / (C_{ham}(W) + C_{spam}(W))$
store the value to a file
Stage2. Filter
For each message M
while (M not end) do
scan message for the next token Ti
query the file for probability of S(Ti)
calculate accumulated message probabilities
S[M] and H[M]
Calculate the overall message filtering indication by:
I[M] = f(S[M] , H[M])
f is a filter dependent function, such as

$$I[M] = \frac{1+S[M]-H[M]}{2}$$

if I[M] > threshold
Email is marked as spam
else
Email is marked as non-spam

We have proposed a Naïve Bayes Algorithm for detecting spam mails using tokens, the product of the individual tokens is used for comparisons. An alternative way to this is to use these tokens against legitimate mails and train the sets to perform accordingly. Also this algorithm can perform faster than other classifier algorithms to give better results.

5. CONCLUSION

Spam detection is a major issue as it saves a lot of time if an automated technique exists to do so.The spam messages are the unwanted messages which the end user clients are receiving in our daily life. Spam mails are nothing it is the advertisement of any company, any kind of virus etcHere we have opted and used naïve Bayes classifier algorithm that aids in spam detection. A serious treat while using email is Spam. We have compared two different classifiers: Support Vector Machine and Naïve Bayes classifier algorithms which help detect spam emails using different approaches.
In this paper we consider factors like scalability, and error rate to choose more genuine approach to escalate the approach of spam detection. In the implementation we use naïve Bayes classifier algorithm to improve the performance. It is a hybrid of rough sets with naïve Bayes, which is implemented in such a way to reduce error rates in spam detection and filtering.

6. REFERENCES

[1] Kecman. Learning and Soft Computing. 2001, The MIT Press.
[2] Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods. 2003, Cambridge University Press.http://www.support-vector.net
[3] M. Sahami et al. A Bayesian Approach to Filtering Junk E-Mail
[4] Enrico Blanzieri and Anton Bryl, "A Survey of Learning Based
[5] Weka. WEKA (Data Mining Software). Available athttp://www.cs.waikato.ac.nz/ml/weka/. 2006
[6] Christina et al. Email Spam Filtering using Supervised Machine Learning Techniques. International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 09, 2010, 3126-3129
[7] Machine Learning Techniques in Spam FilteringKonstantinTretyakov, kt@ut.ee, Institute of Computer Science, University of Tartu, Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004, pp. 60-79.
[8] Priyanka Sao, Pro. KarePrashanthi ,E-mail Spam Classification Using Naïve Bayesian Classifier Khorsi. "An overview of content-based spam filtering techniques", Informatica, 2007
[9] Techniques of Email Spam Filtering," Conference on Email and Anti-Spam., 2008.